Enhancing the Accuracy of AI Image-Editing

Originally published at https://www.unite.ai/enhancing-the-accuracy-of-ai-image-editing/

Although Adobe's Firefly latent diffusion model (LDM) is arguably one of the best currently available, Photoshop users who have tried its generative features will have noticed that it is not able to easily *edit existing images* – instead it completely *substitutes* the user's selected area with imagery based on the user's text prompt (albeit that Firefly is adept at integrating the resulting generated section into the context of the image).

In the current beta version, Photoshop can at least incorporate a reference image as a partial image prompt, which catches Adobe's flagship product up to the kind of functionality that Stable Diffusion users have enjoyed for over two years, thanks to third-party frameworks such as Controlnet:

The current beta of Adobe Photoshop allows for the use of reference images when generating new content inside a selection – though it's a hit-and-miss affair at the moment.

This illustrates an open problem in image synthesis research – the difficulty that diffusion models have in editing existing images without implementing a full-scale 're-imagining' of the selection indicated by the user.

Though this diffusion-based inpaint obeys the user's prompt, it completely reinvents the source subject matter without taking the original image into consideration (except by blending the new generation with the environment). Source: https://arxiv.org/pdf/2502.20376

This problem occurs because LDMs generate images through iterative denoising, where each stage of the process is conditioned on the text prompt supplied by the user. With the text prompt content converted into embedding tokens, and with a hyperscale model such as Stable Diffusion or Flux containing hundreds of thousands (or millions) of near-matching embeddings related to the prompt, the process has a calculated conditional distribution to

aim towards; and each step taken is a step towards this 'conditional distribution target'.

So that's text to image – a scenario where the user 'hopes for the best', since there is no telling exactly what the generation will be like.

Instead, many have sought to use an LDM's powerful generative capacity to edit existing images – but this entails a balancing act between fidelity and flexibility.

When an image is projected into the model's latent space by methods such as DDIM inversion, the goal is to recover the original as closely as possible while still allowing for meaningful edits. The problem is that the more precisely an image is reconstructed, the more the model adheres to its *original* structure, making major modifications difficult.

In common with many other diffusion-based image-editing frameworks proposed in recent years, the Renoise architecture has difficulty making any real change to the image's appearance, with only a perfunctory indication of a bow tie appearing at the base of the cat's throat.

On the other hand, if the process prioritizes editability, the model loosens its grip on the original, making it easier to introduce changes – but at the cost of overall consistency with the source image:

Mission accomplished – but it's a transformation rather than an adjustment, for most AI-based image-editing frameworks.

Since it's a problem that even Adobe's considerable resources are struggling to address, then we can reasonably consider that the challenge is notable, and may not allow of easy solutions, if any.

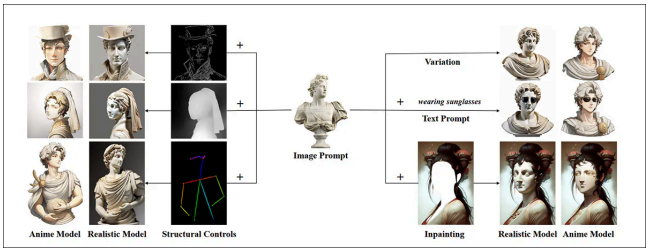
Tight Inversion

Therefore the examples in a new paper released this week caught my attention, as the work offers a worthwhile and noteworthy improvement on the current state-of-the-art in this area, by proving able to apply subtle and refined edits to images projected into the latent space of a model – without the edits either being insignificant or else overwhelming the original content in the source image:

With Tight Inversion applied to existing inversion methods, the source selection is considered in a far more granular way, and the transformations conform to the original material instead of overwriting it.

LDM hobbyists and practitioners may recognize this kind of result, since much of it can be created in a complex workflow using external systems such as Controlnet and IP-Adapter.

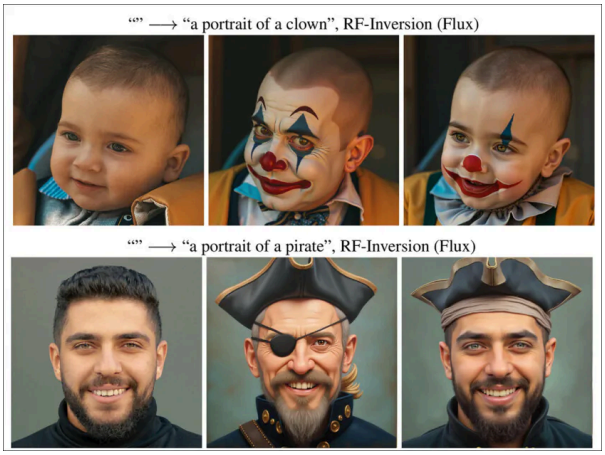
In fact the new method – dubbed *Tight Inversion* – does indeed leverage IP-Adapter, along with a dedicated face-based model, for human depictions.



From the original 2023 IP-Adapter paper, examples of crafting apposite edits to the source material. Source: https://arxiv.org/pdf/2308.06721

The signal achievement of Tight Inversion, then, is to have proceduralized complex techniques into a single drop-in plug-in modality that can be applied to existing systems, including many of the most popular LDM distributions.

Naturally, this means that Tight Inversion (TI), like the adjunct systems that it leverages, uses the source image as a conditioning factor for its own edited version, instead of relying solely on accurate text prompts:



Further examples of Tight Inversion's ability to apply truly blended edits to source material.

Though the authors' concede that their approach is not free from the traditional and ongoing tension between fidelity and editability in diffusion-based image editing techniques, they report state-of-the-art results when injecting TI into existing systems, vs. the baseline performance.

The new work is titled *Tight Inversion: Image-Conditioned Inversion for Real Image Editing*, and comes from five researchers across Tel Aviv University and Snap Research.

Method

Initially a Large Language Model (LLM) is used to generate a set of varied text prompts from which an image is generated. Then the aforementioned DDIM inversion is applied to each image *with three text conditions*: the text prompt used to generate the image; a shortened version of the same; and a null (empty) prompt.

With the inverted noise returned from these processes, the images are again regenerated with the same condition, and without classifier-free guidance (CFG).

Prompt	$L_2 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
Empty	58.972	25.107	0.756	0.264
Short	38.937	28.807	0.858	0.126
Full	21.944	32.526	0.929	0.040
Image prompt	20.497	32.903	0.932	0.035

DDIM inversion scores across various metrics with varying prompt settings.

As we can see from the graph above, the scores across various metrics are improved with increased text length. The metrics used were Peak Signal-to-Noise Ratio (PSNR); L2 distance; Structural Similarity Index (SSIM); and Learned Perceptual Image Patch Similarity (LPIPS).

Image-Conscious

Effectively Tight Inversion changes how a host diffusion model edits real images by conditioning the inversion process on the image itself rather than relying only on text.

Normally, inverting an image into a diffusion model's noise space requires estimating the starting noise that, when denoised, reconstructs the input. Standard methods use a text prompt to guide this process; but an imperfect prompt can lead to errors, losing details or altering structures.

Tight Inversion instead uses IP Adapter to feed visual information into the model, so that it reconstructs the image with greater accuracy, converting the source images into conditioning tokens, and projecting them into the inversion pipeline.

These parameters are editable: increasing the influence of the source image makes the reconstruction nearly perfect, while reducing it allows for more creative changes. This makes Tight Inversion useful for both subtle modifications, such as changing a shirt color, or more significant edits, such as swapping out objects – without the common side-effects of other inversion methods, such as the loss of fine details or unexpected aberrations in the background content.

The authors state:

> *'We note that Tight Inversion can be easily integrated with previous inversion methods (e.g., Edit Friendly DDPM, ReNoise) by [switching the native diffusion core for the IP Adapter altered model], [and] tight Inversion consistently improves such methods in terms of both reconstruction and editability.'*

Data and Tests

The researchers evaluated TI on its capacity to reconstruct and to edit real world source images. All experiments used Stable Diffusion XL with a DDIM scheduler as outlined in the original Stable Diffusion paper; and all tests used 50 denoising steps at a default guidance scale of 7.5.

For image conditioning, IP-Adapter-plus sdxl vit-h was used. For few-step tests, the researchers used SDXL-Turbo with a Euler scheduler, and also conducted experiments with FLUX.1-dev, conditioning the model in the latter case on PuLID-Flux, using RF-Inversion at 28 steps.

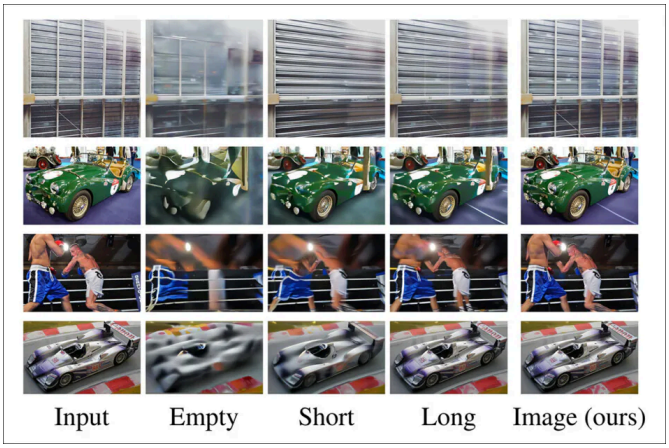
PuIID was used solely in cases featuring human faces, since this is the domain that PuIID was trained to address – and while it's noteworthy that a specialized sub-system is used for this one possible prompt type, our inordinate interest in generating human faces suggests that relying solely on the broader weights of a foundation model such as Stable Diffusion may not be adequate to the standards we demand for this particular task.

Reconstruction tests were performed for qualitative and quantitative evaluation. In the image below, we see qualitative examples for DDIM inversion:

Qualitative results for DDIM inversion. Each row shows a highly detailed image alongside its reconstructed versions, with each step using progressively more precise conditions during inversion and denoising. As the conditioning becomes more accurate, the reconstruction quality improves. The rightmost column demonstrates the best results, where the original image itself is used as the condition, achieving the highest fidelity. CFG was not used at any stage. Please refer to the source document for better resolution and detail.

The paper states:

> *'These examples highlight that conditioning the inversion process on an image significantly improves reconstruction in highly detailed regions.*
>
> *'Notably, in the third example of [the image below], our method successfully reconstructs the tattoo on the back of the right boxer. Furthermore, the boxer's leg pose is more accurately preserved, and the tattoo on the leg becomes visible.'*

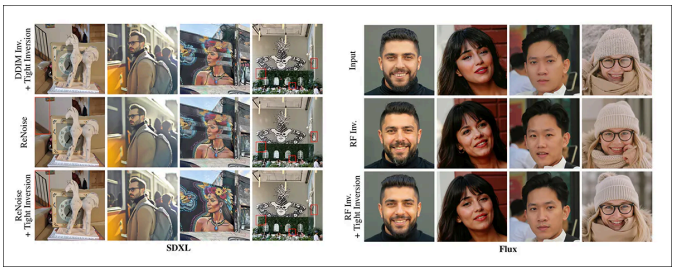


Further qualitative results for DDIM inversion. Descriptive conditions improve DDIM inversion, with image conditioning outperforming text, especially on complex images.

The authors also tested Tight Inversion as a drop-in module for existing systems, pitting the modified versions against their baseline performance.

The three systems tested were the aforementioned DDIM Inversion and RF-Inversion; and also ReNoise, which shares some authorship with the paper under discussion here. Since DDIM results have no difficulty in obtaining 100% reconstruction, the researchers focused only on editability.

(The qualitative result images are formatted in a way that is difficult to reproduce here, so we refer the reader to the source PDF for fuller coverage and better resolution, notwithstanding that some selections are featured below)



Left, qualitative reconstruction results for Tight Inversion with SDXL. Right, reconstruction with Flux. The layout of these results in the published work makes it difficult to reproduce here, so please refer to the source PDF for a true impression of the differences obtained.

Here the authors comment:

> *'As illustrated, integrating Tight Inversion with existing methods consistently improves reconstruction. For [example,] our method accurately reconstructs the handrail in the leftmost example and the man with the blue shirt in the rightmost example [in figure 5 of the paper].'*

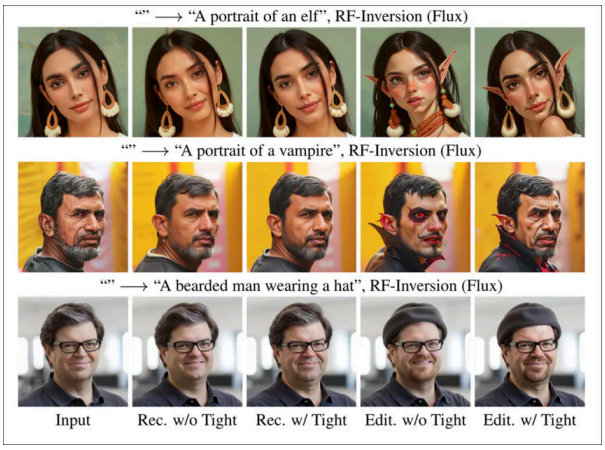
The authors also tested the system quantitatively. In line with prior works, they used the validation set of MS-COCO, and note that the results (illustrated below) improved reconstruction across all metrics for all the methods.

Method	$L_2 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
DDIM Inversion	50.5897	25.3404	0.7699	0.1485
DDIM Inversion + Ours	42.8394	26.9030	0.7981	0.1055
ReNoise	42.9509	27.1584	0.7928	0.1179
ReNoise + Ours	37.8595	28.0413	0.8134	0.0877

Comparing the metrics for performance of the systems with and without Tight Inversion.

Next, the authors tested the system's ability to *edit* photos, pitting it against baseline versions of prior approaches prompt2prompt; Edit Friendly DDPM; LED-ITS++; and RF-Inversion.

Show below are a selection of the paper's qualitative results for SDXL and Flux (and we refer the reader to the rather compressed layout of the original paper for further examples).

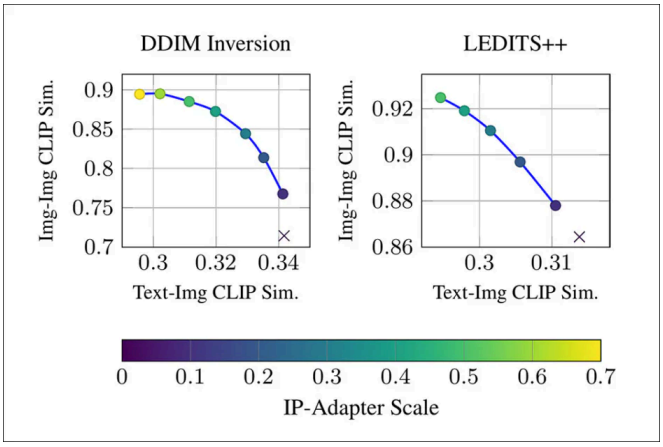


Selections from the sprawling qualitative results (rather confusingly) spread throughout the paper. We refer the reader to the source PDF for improved resolution and meaningful clarity.

The authors contend that Tight Inversion consistently outperforms existing inversion techniques by striking a better balance between reconstruction and editability. Standard methods such as DDIM inversion and ReNoise can recover an image well, the paper states that they often struggle to preserve fine details when edits are applied.

By contrast, Tight Inversion leverages image conditioning to anchor the model's output more closely to the original, preventing unwanted distortions. The authors contend that even when competing approaches produce reconstructions that *appear* accurate, the introduction of edits often leads to artifacts or structural inconsistencies, and that Tight Inversion mitigates these issues.

Finally, quantitative results were obtained by evaluating Tight Inversion against the MagicBrush benchmark, using DDIM inversion and LEDITS++, measured with CLIP Sim.



Quantitative comparisons of Tight Inversion against the MagicBrush benchmark.

The authors conclude:

> *'In both graphs the tradeoff between image preservation and adherence to the target edit is clearly [observed]. Tight Inversion provides better control on this tradeoff, and better preserves the input image while still aligning with the edit [prompt].*
>
> *'Note, that a CLIP similarity of above 0.3 between an image and a text prompt indicates plausible alignment between the image and the prompt.'*

Conclusion

Though it does not represent a 'breakthrough' in one of the thorniest challenges in LDM-based image synthesis, Tight Inversion consolidates a number of burdensome ancillary approaches into a unified method of AI-based image editing.

Although the tension between editability and fidelity is not gone under this method, it is notably reduced, according to the results presented. Considering that the central challenge this work addresses may prove ultimately intractable if dealt with on its own terms (rather than looking beyond LDM-based architectures in future systems), Tight Inversion represents a welcome incremental improvement in the state-of-the-art.

First published Friday, February 28, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More